AJAY KADIYALA - Data Engineer

Follow me Here:

LinkedIn:
https://www.linkedin.com/in/ajay026/

Data Geeks Community:

https://lnkd.in/gU5NkCqi

500+

DATA ENGINEERING INTERVIEW QUESTOINS

1. What is Hadoop MapReduce?

A.) For processing large datasets in parallel across hadoop cluster, hadoop mapReduce framework is used.

2. What are the difference between relational database and HDFS?

A.) There are 6 major categories we can define RDMBS and HDFS. They are

a. Data Types

b. processing

c. Schema on read Vs Write

d. Read/write speed

e. cost

Best fit use case

RDBMS	HDFS
1. In RDBMS it relies on structured data and schema is always known.	any kind of data can be stored into Hadoop. i.e structured, unstructured, semi-structured.
2. Rdbms provides limited or no processing capabilites.	hadoop allows us to process the data which is distributed across the cluster in a parallel fashion.
3. Rdbms is based schema on write.	Hadoop follows schema on read policy.

4. in rdbms reads are fast because the schema is already known. writes are fast in hadoop because no schema validation happens during hdfs write.

5. Licensed software, therefore, need to pay for software. Hadoop is open source framework, hence no need to pay for software.

6. Rdbms is used for OLTP(online transactional processing) system. Hadoop is used for data discovery, data analytics or OLAP System.

3. Explain Bigdata and explain 5v's of bigdata?

A.) Bigdata is a term for collection of large and complex datasets, that makes it to difficult processing using relational database management tools or traditional data processing applications. It is difficult capture, visualize, curate, store, search, share, transfer and analyze bigdata.

IBM has defined bigdata with 5v's they are.

a. Volume

b. velocity

c. variety

d. veracity

e. Value: It is good to have access to bigdata but unless we turn into value it is useless. which means using bigdata adding benifits to the organizations, and are they seen ROI using Bigdata.

4. What is Hadoop and its components?

A.) When bigdata emerged as a problem, Hadoop evolved as Solution to bigdata. Apache hadoop is a framework which provides us various services or tools, to store and process the bigdata.

It helps in analyzing bigdata and making bussiness decisions out of it, which cannot be done using traditional systems.

Main components in Hadoop are

a. Storage(Namenode, DataNode)

b. Processing framework yarn (resource manager, Node Manager)

5. What are HDFS and Yarn?

A.) HDFS (Hadoop Distributed File System) is the storage unit of Hadoop. It is responsible for storing different kinds of data as blocks in a distributed environment.

It follows master and slave topology.

NameNode: NameNode is the master node in the distributed environment and it maintains the metadata information for the blocks of data stored in HDFS like block location, replication factors etc.

DataNode: DataNodes are the slave nodes, which are responsible for storing data in the HDFS. NameNode manages all the DataNodes.

YARN (Yet Another Resource Negotiator) is the processing framework in Hadoop, which manages resources and provides an execution environment to the processes.

ResourceManager: It receives the processing requests, and then passes the parts of requests to corresponding NodeManagers accordingly, where the actual processing takes place. It allocates resources to applications based on the needs.

NodeManager: NodeManager is installed on every DataNode and it is responsible for the execution of the task on every single DataNode.

6. Tell me about various Hadoop Daemons and their roles in hadoop cluster?

A.) Generally approach this question by first explaining the HDFS daemons i.e. NameNode, DataNode and Secondary NameNode, and then moving on to the YARN daemons i.e. ResorceManager and NodeManager, and lastly explaining the JobHistoryServer.

JobHistoryServer: It maintains information about MapReduce jobs after the Application Master terminates.

7. Compare HDFS with Network attached servive(NAS)?

A.) Network-attached storage (NAS) is a file-level computer data storage server connected to a computer network providing data access to a heterogeneous group of clients. NAS can either be a

hardware or software which provides services for storing and accessing files.

Whereas Hadoop Distributed File System (HDFS) is a distributed filesystem to store data using commodity hardware.

In HDFS Data Blocks are distributed across all the machines in a cluster.

Whereas in NAS data is stored on a dedicated hardware.

8. List the difference between Hadoop 1.0 vs Hadoop 2.0?

A.) To answer this we need to highlight 2 important features that are a. PassiveNode b. Processing.

In Hadoop 1.x, “NameNode” is the single point of failure. In Hadoop 2.x, we have Active and Passive “NameNodes”. If the active “NameNode” fails, the passive “NameNode” takes charge. Because of this, high availability can be achieved in Hadoop 2.x.

Also, in Hadoop 2.x, YARN provides a central resource manager. With YARN, you can now run multiple applications in Hadoop, all sharing a common resource. MRV2 is a particular type of distributed application that runs the MapReduce framework on top of YARN. Other tools can also perform data processing via YARN, which was a problem in Hadoop 1.x.

9. What are active and Passive Namenodes?

A.) In a High availability architecture, there are 2 namenodes. i.e.

a. Active “NameNode” is the “NameNode” which works and runs in the cluster.

b. Passive “NameNode” is a standby “NameNode”, which has similar data as active “NameNode”.

When the active “NameNode” fails, the passive “NameNode” replaces the active “NameNode” in the cluster.

10. Why does one remove or add datanodes freaquently?

A.) most attractive features of the Hadoop framework is its utilization of commodity hardware. However, this leads to frequent “DataNode” crashes in a Hadoop cluster.

Another striking feature of Hadoop Framework is the ease of scale in accordance with the rapid growth in data volume. this is why hadoop admins work is to commission or decommission nodes in hadoop cluster.

11.) what happens when two clients tries to access same file in Hdfs?

A.) When first client request for file or data hdfs provides access to write, but when second client request it rejects by saying already another client accessing it.

12. How does nameNOde tackles data node failures?

A.) NameNode periodically receives a Heartbeat (signal) from each of the DataNode in the cluster, which implies DataNode is functioning properly.

13. What will you do when NameNode is down?

A.) Use the file system metadata replica (FsImage) to start a new NameNode.

Then, configure the DataNodes and clients so that they can acknowledge this new NameNode, that is started.

14. What is checkpoint?

A.) Checkpointing" is a process that takes an FsImage, edit log and compacts them into a new FsImage.

Thus, instead of replaying an edit log, the NameNode can load the final in-memory state directly from the FsImage.

15. what is Hdfs fault tolerant?

A.) When data stored in over hdfs, namenode replicates the data in server datanode, which has default value of 3.

if any DataNode fails Namenode automatically copies data to another datanode to makesure data is fault tolerant.

16. can NameNode and dataNode are commodity hardware?

A.) No.. beacuse Namenode built in with high memory space, and with higher quality software.

but datanode built in with cheaper hardware.

17. Why do we use Hdfs for files with large data sets but not when there are lot of small files?

A.) NameNode stores the metadata information regarding the file system in the RAM.

Therefore, the amount of memory produces a limit to the number of files in my HDFS file system. In other words, too many files will lead to the generation of too much metadata.

And, storing these metadata in the RAM will become a challenge. hence hdfs is only works with large datsets instead large no.of small.

18. How do you define block, and what is the default block size?

A.) Block are nothing but smallest continuous locations in harddrive where data is stored.

default block size of hadoop 1 is 64 mb

and hadoop 2 is 128 mb.

19. How do you define Rack awareness in hadoop?

A.) Rack Awareness is the algorithm in which the “NameNode” decides how blocks and their replicas are placed, based on rack

definitions to minimize network traffic between “DataNodes” within the same rack.

20. What is the difference between hdfs block, and input split?

A.) The “HDFS Block” is the physical division of the data while “Input Split” is the logical division of the data.

Hdfs block divides data into blocks to store the blocks together processing, where Input split Divides the data into the input split and assign it to the mapper function for processing.

21. Name of three modes which hadoop can run?

A.) Stanalone mode

pseudo- distribution mode

fully distributed mode

29. What do you know about SequenceFileFormat?

A.) “SequenceFileInputFormat” is an input format for reading within sequence files.

It is a specific compressed binary file format which is optimized for passing the data between the outputs of one “MapReduce” job to the input of some other “MapReduce” job.

30. What is Hive?

A.) Apache Hive is a data warehouse system built on top of Hadoop and is used for analyzing structured and semi-structured data developed by Facebook.

Hive abstracts the complexity of Hadoop MapReduce.

31. what is Serde in Hive?

A.) The “SerDe” interface allows you to instruct “Hive” about how a record should be processed.

A “SerDe” is a combination of a “Serializer” and a “Deserializer”.

“Hive” uses “SerDe” (and “FileFormat”) to read and write the table’s row.

31. can the default hive metastore used by multiple users at the same time?

A.) “Derby database” is the default “Hive Metastore”.

Multiple users (processes) cannot access it at the same time.

It is mainly used to perform unit tests.

32. what is the default location for hive to store in table data?

A.) The default location where Hive stores table data is inside HDFS in /user/hive/warehouse.

33. What is Apache Hbase?

A.) HBase is an open source, multidimensional, distributed, scalable and a NoSQL database written in Java.

HBase runs on top of HDFS (Hadoop Distributed File System) and provides BigTable (Google) like capabilities to Hadoop.

It is designed to provide a fault-tolerant way of storing the large collection of sparse data sets.

HBase achieves high throughput and low latency by providing faster Read/Write Access on huge datasets.

34. What are the components of apache Hbase?

A.) HBase has three major components, i.e. HMaster Server, HBase RegionServer and Zookeeper.

Region Server: A table can be divided into several regions. A group of regions is served to the clients by a Region Server.

HMaster: It coordinates and manages the Region Server (similar as NameNode manages DataNode in HDFS).

ZooKeeper: Zookeeper acts like as a coordinator inside HBase distributed environment.

It helps in maintaining server state inside the cluster by communicating through sessions.

35. what are the components of Region server?

A.) WAL: Write Ahead Log (WAL) is a file attached to every Region Server inside the distributed environment. The WAL stores the new data that hasn't been persisted or committed to the permanent storage.

Block Cache: Block Cache resides in the top of Region Server. It stores the frequently read data in the memory.

MemStore: It is the write cache. It stores all the incoming data before committing it to the disk or permanent memory. There is one MemStore for each column family in a region.

HFile: HFile is stored in HDFS. It stores the actual cells on the disk.

36. What is the difference between Hbase and Relation database?

A.) HBase is an open source, multidimensional, distributed, scalable and a NoSQL database written in Java.

Hbase	Relational Database
1. It is schema-less	It is schema-based database.
2. It is column-oriented data store	It is row-oriented data store.
3. It is used to store de-normalized data	It is used to store normalized data.
4. Automated partitioning is done is HBase	There is no such provision or built-in support for partitioning.

35. What is Apache Spark?

A.) Apache Spark is a framework for real-time data analytics in a distributed computing environment.

It executes in-memory computations to increase the speed of data processing.

It is 100x faster than MapReduce for large-scale data processing by exploiting in-memory computations and other optimizations.

36. can you build Spark with any particular Hadoop version?

A.) yes. spark can be built with any version of Hadoop.

37. What is RDD?

A.) RDD is the acronym for Resilient Distribution Datasets – a fault-tolerant collection of operational elements that run parallel.

The partitioned data in RDD are immutable and distributed, which is a key component of Apache Spark.

38. Are Hadoop and Bigdata are co related?

A.) Big Data is an asset, while Hadoop is an open-source software program, which accomplishes a set of goals and objectives to deal with that asset.

Hadoop is used to process, store, and analyze complex unstructured data sets through specific proprietary algorithms and methods to derive actionable insights.

So yes, they are related but are not alike.

39. why is Hadoop used in bigdata analytics?

A.) Hadoop allows running many exploratory data analysis tasks on full datasets, without sampling. Features that make Hadoop an essential requirement for Big Data are –

Data collection

Storage

Processing

Runs independently.

40. Name of some of the important tools used for data analytics?

A.) The important Big Data analytics tools are –

NodeXL

KNIME

Tableau

Solver

OpenRefine

Rattle GUI

Qlikview.

41. what is FSCK?

A.) FSCK or File System Check is a command used by HDFS.

It checks if any file is corrupt,or if there are some missing blocks for a file. FSCK generates a summary report, which lists the overall health of the file system.

42. what are the different core methods of Reducer?

A.) There are three core methods of a reducer-

setup() – It helps to configure parameters like heap size, distributed cache, and input data size.

reduce() – Also known as once per key with the concerned reduce task. It is the heart of the reducer.

cleanup() – It is a process to clean up all the temporary files at the end of a reducer task.

43. what are the most common Input fileformats in Hadoop?

A.) The most common input formats in Hadoop are –

Key-value input format

Sequence file input format

Text input format.

44. what are the different fileformats that can be used in Hadoop?

A.) File formats used with Hadoop, include –

CSV

JSON

Columnar

Sequence files

AVRO

Parquet file.

45. what is commodity hardware?

A.) Commodity hardware is the basic hardware resource required to run the Apache Hadoop framework.

It is a common term used for affordable devices, usually compatible with other such devices.

46. what do you mean by logistic regression?

A.) Also known as the logit model, logistic regression is a technique to predict the binary result from a linear amalgamation of predictor variables.

47. Name the port number for namenode, task tracker, job tracker?

A.) NameNode – Port 50070

Task Tracker – Port 50060

Job Tracker – Port 50030.

48. Name the most popular data management tools that used with edge nodes in hadoop?

A.) The most commonly used data management tools that work with Edge Nodes in Hadoop are –

Oozie

Ambari

Pig

Flume.

49. what is block in Hadoop distributed file system?

A.) When the file is stored in HDFS, all file system breaks down into a set of blocks.

50. what is the functionality of jps command?

A.) The ‘jps’ command enables us to check if the Hadoop daemons like namenode, datanode, resourcemanager, nodemanager, etc. are running on the machine.

51. what types of biases can happen through sampling?

A.) Three types of biases can happen through sampling, which are –

Survivorship bias

Selection bias

Under coverage bias.

52. what is the difference between Sqoop and distcp?

A.) DistCP is used for transferring data between clusters, while Sqoop is used for transferring data between Hadoop and RDBMS, only.

53.How much data is enough to get a valid outcome?

A.) The amount of data required depends on the methods you use to have an excellent chance of obtaining vital results.

54. Is Hadoop is different from other parallel computing systems? How?

A.) Yes, it is. Hadoop is a distributed file system. It allows us to store and manage large amounts of data in a cloud of machines, managing data redundancy.

The main benefit of this is that since the data is stored in multiple nodes, it is better to process it in a distributed way. Each node is able to process the data stored on it instead of wasting time moving the data across the network.

In contrast, in a relational database computing system, we can query data in real-time, but it is not efficient to store data in tables, records, and columns when the data is huge.

Hadoop also provides a schema for building a column database with Hadoop HBase for run-time queries on rows.

55. What is BackUp Node?

A.) Backup Node is an extended checkpoint node for performing checkpointing and supporting the online streaming of file system edits.

Its functionality is similar to Checkpoint, and it forces synchronization with NameNode.

56. what are the common data challenges?

A.) The most common data challenges are –

Ensuring data integrity

Achieving a 360-degree view

Safeguarding user privacy

Taking the right business action with real-time resonance.

57. How do you overcome above mentioned data challenges?

A.) Data challenges can be overcome by –

Adopting data management tools that provide a clear view of data assessment

Using tools to remove any low-quality data

Auditing data from time to time to ensure user privacy is safeguarded

Using AI-powered tools, or software as a service (SaaS) products to combine datasets and make them usable.

58. What is the hierarachical Clustering algorithm?

A.) The hierarchical grouping algorithm is the one that combines and divides the groups that already exist.

58. what is K- Mean clustering?

A.) K mean clustering is a method of vector quantization.

59. can you mention the crieteria for good data model?

A.) A good data model –

It should be easily consumed

Large data changes should be scalable

Should offer predictable performances

Should adapt to changes in requirements.

60. Name the different commands for starting up and shutting down the hadoop daemons?

A.) To start all the daemons:

./sbin/start-all.sh

To shut down all the daemons:

./sbin/stop-all.sh

61. Talk about the different tombstone markers used for deletion purpose in Hbase?

A.) There are three main tombstone markers used for deletion in HBase. They are-

Family Delete Marker – For marking all the columns of a column family.

Version Delete Marker – For marking a single version of a single column.

Column Delete Marker – For marking all the versions of a single column.

62. How can bigdata add value to bussinesses?

A.) ig Data Analytics helps businesses to transform raw data into meaningful and actionable insights that can shape their business strategies.

The most important contribution of Big Data to business is data-driven business decisions.

63. How do you deploy bigdata solution?

A.) we can deploy bigdata in 3 stages. they are

Data Ingestion: begin by collecting data from multiple sources, be it social media platforms, log files, business documents, anything relevant to your business.

Data can either be extracted through real-time streaming or in batch jobs.

Data storage: Once the data extracted, it can be stored in Hbase, or Hdfs.

While HDFS storage is perfect for sequential access, HBase is ideal for random read/write access.

Data processing: Usually, data processing is done via frameworks like Hadoop, Spark, MapReduce, Flink, and Pig, to name a few.

64. List the different file permissions in hdfs files or directory levels?

A.) There are three user levels in HDFS – Owner, Group, and Others. For each of the user levels, there are three available permissions:

read (r)

write (w)

execute(x).

65. Elaborate on the process that overwrite the replication factor in Hdfs?

A.) In HDFS, there are two ways to overwrite the replication factors – on file basis and on directory basis.

66. Explain overFitting?

A.) Overfitting refers to a modeling error that occurs when a function is tightly fit (influenced) by a limited set of data points.

67. what is feature selection?

A.) Feature selection refers to the process of extracting only the required features from a specific dataset.

When data is extracted from disparate sources.

Feature selection can be done via three techniques.

a. filters method

b. wrappers method

c. Embedded method.

68. Define OUtliers?

A.) outliers are the values that are far removed from the group, they do not belong to any specific cluster or group in the dataset.

The presence of outliers usually affects the behavior of the model.

Here are six outlier detection methods:

1. Extreme value analysis

2. probabilistic analysis

3. linear models

4. information-theoretic models

5. High-dimensional outlier detection.

70. How can you handle missing values in Hadoop?

A.) there are different ways to estimate the missing values.

These include regression, multiple data imputation, listwise/pairwise deletion, maximum likelihood estimation, and approximate Bayesian bootstrap.

MapReduce Interview Questions:

1. Compare MapReduce and SPark?

A.) there are 4 crieteria to be followed to compare MR with spark. they are.

1. processing speeds

2. standalone mode

3. Ease of use

4. versatility

MapReduce	Spark
1. processing speed is good	It is execeptional
2. standalone mode needs hadoop	it can work independently
3. it needs extensive java program	API for python & scala& java
4. It is optimized real time machine-learning applications.	not optimized real time & mL applications.

2. what is MapReduce?

A.) It is a framework/a programming model that is used for processing large data sets over a cluster of computers using parallel programming.

3. State the reason why we can't perform aggregation in mapper? why do we need reducer for this?

A.) We cannot perform “aggregation” (addition) in mapper because sorting does not occur in the “mapper” function, as sorting occurs only on reducer.

During “aggregation”, we need the output of all the mapper functions which may not be possible to collect in the map phase as mappers may be running on the different machine.

4. What is the recordReader in Hadoop?

A.) The “RecordReader” class loads the data from its source and converts it into (key, value) pairs suitable for reading by the “Mapper” task.

5. Explain Distributed cache in MapReduce Framework?

A.) Distributed Cache is a dedicated service of the Hadoop MapReduce framework, which is used to cache the files whenever required by the applications.

This can cache read-only text files, archives, jar files, among others, which can be accessed and read later on each data node where map/reduce tasks are running.

6. How do reducers communicate with each other?

A.) The “MapReduce” programming model does not allow “reducers” to communicate with each other.

6. What does mapReduce partitioner do?

A.) A “MapReduce Partitioner” makes sure that all the values of a single key go to the same “reducer”, thus allowing even distribution of the map output over the “reducers”.

It redirects the “mapper” output to the “reducer” by determining which “reducer” is responsible for the particular key.

7. How will you write custom partitioner?

A.) custom partitioner can be written in following ways.

Create a new class that extends Partitioner Class

Override method – getPartition, in the wrapper that runs in the MapReduce.

Add the custom partitioner to the job by using method set Partitioner.

8. What is combiner?

A.) A “Combiner” is a mini “reducer” that performs the local “reduce” task.

It receives the input from the “mapper” on a particular “node” and sends the output to the “reducer”.

9. what are main components of MapReduce?

A.) Main Driver Class: providing job configuration parameters

Mapper Class: must extend org.apache.hadoop.mapreduce.Mapper class and performs execution of map() method

Reducer Class: must extend org.apache.hadoop.mapreduce.Reducer class.

10. What is Shuffling and Sorting in MapReduce?

A.) Shuffling and Sorting are two major processes operating simultaneously during the working of mapper and reducer.

The process of transferring data from Mapper to reducer is Shuffling.

In MapReduce, the output key-value pairs between the map and reduce phases (after the mapper) are automatically sorted before moving to the Reducer.

11. What is identity mapper and Chain mapper?

A.) Identity Mapper is the default Mapper class provided by Hadoop.

It only writes the input data into output and do not perform and computations and calculations on the input data.

Chain mapper: Chain Mapper is the implementation of simple Mapper class through chain operations across a set of Mapper classes, within a single map task.

In this, the output from the first mapper becomes the input for second mapper

12. What main configuration parameters are specified in Mapreduce?

A.) following configuration parameters to perform the map and reduce jobs:

The input location of the job in HDFs.

The output location of the job in HDFS.

The input's and output's format.

The classes containing map and reduce functions, respectively.

The .jar file for mapper, reducer and driver classes.

13. Name Job control options specified by mapreduce?

A.) Since this framework supports chained operations wherein an input of one map job serves as the output for other.

The various job control options are:

Job.submit() : to submit the job to the cluster and immediately return

Job.waitforCompletion(boolean) : to submit the job to the cluster and wait for its completion.

14. What is inputFormat in hadoop?

A.) inputformat defines the input specifications for a job.it peforms following instructions.

1. validates input-specifications of job.

2. Split the input file(s) into logical instances called InputSplit.

3. Provides implementation of RecordReader to extract input records from the above instances for further Mapper processing.

15. What is the difference between Hdfs block and inputsplit?

A.) An HDFS block splits data into physical divisions while InputSplit in MapReduce splits input files logically.

16. what is the text inputformat?

A.) TextInputFormat, files are broken into lines, wherein key is position in the file and value refers to the line of text.

Programmers can write their own InputFormat.

17. what is role of job Tracker?

A.) The primary function of the JobTracker is resource management, which essentially means managing the TaskTrackers.

Apart from this, JobTracker also tracks resource availability and handles task life cycle management.

18. Explian jobconf in mapreduce?

A.) It is a primary interface to define a map-reduce job in the Hadoop for job execution.

JobConf specifies mapper, Combiner, partitioner, Reducer,InputFormat , OutputFormat implementations

19. what is output committer?

A.) OutPutCommitter describes the commit of MapReduce task.

FileOutputCommitter is the default available class available for OutputCommitter in MapReduce.

20. what is map in Hadoop?

A.) In Hadoop, a map is a phase in HDFS query solving.

A map reads data from an input location, and outputs a key value pair according to the input type.

21. what is reducer in hadoop?

A.) In Hadoop, a reducer collects the output generated by the mapper, processes it, and creates a final output of its own.

22. what are the parameters of mappers and reducers?

A.) The four parameters for mappers are:

LongWritable (input)

text (input)

text (intermediate output)

IntWritable (intermediate output)

The four parameters for reducers are:

Text (intermediate output)

IntWritable (intermediate output)

Text (final output)

IntWritable (final output)

23. What is partitioning?

A.) Partitioning is a process to identify the reducer instance which would be used to supply the mappers output.

Before mapper emits the data (Key Value) pair to reducer, mapper identify the reducer as an recipient of mapper output.

24. what is mapreduce used for-by company?

A.) construction of index for google search: The process of constructing a positional or nonpositional index is called index construction or indexing.

article clustering for google news: For article clustering, the pages are first classified according to whether they are needed for clustering.

statistical machine transalation.

25 what are the mapreduce design goals?

A.) scalability to large data volumes

cost-efficiency.

26. what are the challenges of Mapreduce?

A.) cheap node fails, specially if having many.

a commodity network is equao to implies.

programming distributed systems are hard.

27. what is the mapreduce programming model?

A.) MapReduce programming model is based on a concept called key-value records.

It also provides paradigms for parallel data processing.

28. what are the mapreduce execution details?

A.) In the case of MapReduce execution, a single master controls job execution on multiple slaves.

29. Mention benifits of Mapreduce?

A,) Highly scalable

cost-effective

secure.

30. is the renaming the output file possible?

A.) yes, the implementation of multiple format output class makes it possible to rename the output file.

Apache Sqoop Interview Questions:

1. Mention the best features of Apache Sqoop?

A.) Apache sqoop is a tool in hadoop ecosystem have several advantages. i.e.

Parallel import/export

Connectors for all major RDBMS Databases

Import results of SQL query

Incremental Load

Full Load

Kerberos Security Integration

Load data directly into Hive / HBase

Compression

Support for Accumulo

2. How can you import large objects like BLOB and CLOB in sqoop?

A.) The direct import function is not supported by Sqoop in case of CLOB and BLOB objects. Hence, if you have to import large purposes, you can use JDBC based imports.

This can be done without introducing the direct argument of the import utility.

3. What is default database of Apache sqoop?

A.) The default database of apache sqoop is MySQL.

4. Describe the process of executing free-form SQL query to import rows?

A.) To achieve a free-form SQL query, you have to use the –m1 option. This would create only one Mapreduce task.

This would then import the rows directly.

5. Describe the importance of using compress-codec parameter?

A.) The –compress-codec parameter can be used to get the export file of the Sqoop import in the required formats.

6. What is the significance of Sqoop eval tool?

A.) Sqoop eval can be against the database as it can preview the results that are displayed on the console. Interestingly, with the help of the Eval tool, you would be well aware of the fact that the desired data can be imported correctly or not.

7. What is the meaning of free form import in sqoop?

A.) With the use of Sqoop, one can import the relational database query. This can be done using column and table name parameters.

8. Describe the advantage of utilizing --password-file rather than -p option?

A.) The –password-file option is usually used inside the Sqoop script file. On the other hand, the –P option is able to read the standard input along with the column name parameters.

9. Is the JDBC driver fully capable to connect sqoop on the databases?

A.) The JDBC driver is not capable to connect Sqoop on the databases. This is the reason that Sqoop requires both the connector and JDBC driver.

10. what is the meaning of input split in Hadoop?

A.) Input Split is that kind of a function which is associated with splitting the input files into various chunks. These chunks can also assign each split to a mapper in the ongoing process of data correction.

11. Illustrate the utility of --help command in sqoop?

A.) The command in the sqoop can be utilized to list the various available commands.

12. what is Codegen commnad in sqoop?

A.) The Codegen command is associated with the generation of code so that it can appropriately interact with the database records.

13. Describe the procedure involved in executing an incremental data load in sqoop?

A.) the process of performing additional data load is to update the uploaded data. This data is often referred to as delta data. In Sqoop, this delta data can be altered with the use of incremental load command.

14. What is the default file format in order to import data with the utilization of apache sqoop?

A.) Delimiting the text File Format.

15. List all basic sqoop commands along with their properties?

A.) The basic controls in Apache Sqoop along with their uses are:

1. Export: This function helps to export the HDFS directory into a database table

2. List Tables: This function would help the user to list all tables in a particular database.

3. Codegen: This function would help you to generate code so that you can interact with varied types of database records.

4. Create: This function allows a user to import the table definition within the hive of databases.

5. Eval: This function would always help you to assess the SQL statement and display the results.

6. Version: This function would help you to depict the information related to the text of the database.

7. Import all tables: This function would help a user to import all the tables from a database to HDFS.

8. List all the databases: This function would assist a user to create a list of the available databases on a particular server.

16. what are the limitations of importing the RDBMS tables into Hcatlog directly?

A.) In order to import the tables into the Hcatalog in a direct manner, you have to make sure that you are using the –Hcatalog database option. However, in this process, you would face a limitation of importing the tables.

It is in the form of the fact that this option do not supports a plethora of arguments like –direct, –as-Avro file and -export-dir.

17. what is the procedure of updating the rows that have been directly uploaded?

A.) In order to update the existing rows that have been exported, we have to use parameter is in the form of update key

18. What is the significance of sqoop import Mainframe tool?

A.) The Sqoop Import Mainframe tool can also be used to import all the important datasets which lies in a partitioned dataset.

This tool would always help you to make sure that you are importing the right types of data tools and that too in a proper manner.

19. Define Metastore?

A.) It is also known as a shared metadata repository with the help of which the local users can execute and define various types of list tables.

In order to connect to the metastore, you have to make changes to the Sqoop –site.xml.

20. Does sqoop uses MapReduce Function?

A.) Apache Sqoop also uses the Map-Reduce function of Hadoop to obtain data from the relational databases.

During the process of importing data, Sqoop controls the mappers and their numbers.

21. Compare Sqoop and Flume?

A.) criteria Sqoop Flume

Application Importing data from RDBMS Moving bulk streaming data into HDFS

Architecture Connector-connecting to respective data
Agent – fetching of the right data

Loading of data Event driven Not event driven

22. How can we import data from particular row or column?

A.) Sqoop allows to Export and Import the data from the data table based on the where clause.

23. Role of JDBC driver in sqoop setup?

A.) Sqoop needs a connector to connect the different relational databases. Almost all Database vendors make a JDBC connector available specific to that Database, Sqoop needs a JDBC driver of the database for interaction.

No, Sqoop needs JDBC and a connector to connect a database.

24. Using Sqoop command how can we control the number of Mappers?

A.) We can control the number of mappers by executing the parameter –num-mapers in sqoop command.

25. What is the purpose of Sqoop-merge?

A.) This tool combines 2 datasets where entries in one dataset overwrite entries of an older dataset preserving only the new version of the records between both the data sets.

26. Explain the Saved Job process in Sqoop?

A.) Sqoop allows us to define saved jobs which make this process simple. A saved job records the configuration information required to execute a Sqoop command at a later time.

sqoop-job tool describes how to create and work with saved jobs.

27. Sqoop is Which type of tool and main use of Sqoop?

A.) And Sqoop is a data transfer tool.

The main use of Sqoop is to import and export the large amount of data from RDBMS to HDFS and vice versa.

28. I am getting connection failure exception during connecting to Mysql through Sqoop, what is the root cause and fix for this error scenario?

A.) This will happen when there is lack of permissions to access our Mysql database over the network.

We can try the below command to confirm the connect to Mysql database from aSqoop client machine.

$ mysql –host=MySqlnode> –database=test –user= –password=

We can grant the permissions with below commands.

mysql> GRANT ALL PRIVILEGES ON *.* TO ‘%’@’localhost’;

mysql> GRANT ALL PRIVILEGES ON *.* TO ‘ ’@’localhost’;

29. I am getting java.lang.IllegalArgumentException: during importing tables from oracle database.what might be the root cause and fix for this error scenario?

A.) Sqoop commands are case- sensitive of table names and user names.

By specifying the above two values in UPPER case, it will resolve the issue.

30. Is Sqoop same as to distcp in Hadoop?

A.) No. Because the only the distcp import command is same as Sqoop import command and both the commands submit parallel map-only jobs but both command functions are different.

Distcp is used to copy any type of files from Local filesystem to HDFS and Sqoop is used for transferring the data records between RDBMS and Hadoop eco- system service.

31. I am having around 500 tables in a database. I want to import all the tables from the database except the tables named Table 498, Table 323, and Table 199. How can we do this without having to import the tables one by one?

A.) This can be proficient using the import-all-tables, import command in Sqoop and by specifying the exclude-tables option with it as follows-

sqoop import-all-tables

–connect –username –password –exclude-tables Table498, Table 323, Table 199

32. Explian the significance of using -split-by clause in Sqoop?

A.) split-by is a clause, it is used to specify the columns of the table which are helping to generate splits for data imports during importing the data into the Hadoop cluster.

This clause specifies the columns and helps to improve the performance via greater parallelism

33. If the source data gets updated every now and then, how will you synchronize the data in HDFS that is imported by Sqoop?

A.) By using incremental parameter we can syncronize the data.

and also we can use append, and lastmodified modes to update existing data.

34. When to use target-dir and when to use warehouse-dir in sqoop?

A.) we use –target-dir to specify a particular directory in HDFS.

Whereas we use –warehouse-dir to specify the parent directory of all the sqoop jobs.

35. what is the purpose of validation in sqoop?

A.) In Sqoop, validating the data copied is Validation's main purpose.

Basically, either Sqoop import or Export by comparing the row counts from the source as well as the target post copy.

36. what is accumulo in sqoop?

A.) The Accumulo in sqoop is a sorted, distributed key and value store. It provides robust, extensible data storage and retrieves data.

37. Explain relaxed isolation in sqoop?

A.) This is used to import the data which is read uncommitted for mappers.

The sqoop transfer committed data relational database to the Hadoop file system but with this argument, we can transfer uncommitted data in the isolation level.

38. What are reducers in Sqoop?

A.) The reducer is used for accumulation or aggregation.

In the sqoop there is no reducer because import and export work parallel in sqoop.

39. What is boundary query in sqoop?

A.) The boundary query is used for splitting the value according to id_no of the database table.

To boundary query, we can take a minimum value and maximum value to split the value.

To make split using boundary queries, we need to know all the values in the table.

40. What is sqoop?

A.) The sqoop is an acronym of SQL-TO-HADOOP.

Sqoop is a tool used to transfer to data between Hadoop and relational databases or vice versa.

1. what is spark?

A.) Spark is General purpose, in memory compute engine.

General purpose: it can support any storage, any compute engine

in memory: Spark save storage in memory rather disc in MapReduce.

compute engine: It is plug and play compute engine.

2. Difference between spark & MR?

A.) Performance: Spark was designed to be faster than MapReduce, and by all accounts, it is; in some cases, Spark can be up to 100 times faster than MapReduce.

Operability: Spark is easier to program than MapReduce.

Data Processing: MapReduce and Spark are both great at different types of data processing tasks.

Failure Recovery

Security.

3. Explain the architecture of spark?

A.) Spark Architecture. The Spark follows the master-slave architecture. Its cluster consists of a single master and multiple slaves. The Spark architecture depends upon two abstractions:

Resilient Distributed Dataset (RDD) Directed Acyclic Graph (DAG) Resilient Distributed Datasets (RDD)

In spark there are 2 kinds of operations.

1. Transformations.

2. Actions.

transformations are lazy which means when we execute the below lines, no actual computation has happened but a diagram will be created.

but actions are not.

A DAG is generated when we compute spark statements.

Execution happens when action is encountered before that only entries are made into DAG.

4. What is RDD?

A.) Resilient Distributed Datasets Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark. It is an immutable distributed collection of objects.

Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster.

Resilient: ability to quickly recover from failures.

5. How spark achieves fault tolerance?

A.) Rdd Provides Fault tolerance through lineage graph.

A lineage graph keeps a track of transformations to be executed after action has been called.

6. What is Transformations & action in spark?

A.) In spark there are 2 kinds of operations.

1. Transformations.

2. Actions.

tranformations are lazy which means when we execute the below lines, no actual computation has happend but a diagram will be created.

but actions are not.

if transformations are not lazy, we just need query simple function to large file.

7. Difference between Narrow & wide transformations?

A.) A narrow transformation is one in which a single input partition maps to a single output partition.

example filter,map, flatmap

A wide transformation is a much more expensive operation and is sometimes referred to as a shuffle in Spark.

Ex: ReduceByKey, Groupby

8. what is difference between DAG & Lineage?

A.) DAG:A DAG is generated when we compute spark statements.

Execution happens when action is encountered before that only entries are made into DAG.

Lineage: Rdd Provides Fault tolerance through lineage graph.

A lineage graph keeps a track of transformations to be executed after action has been called.

9. What is partition and how spark Partitions the data?

A.) A Partition in simple terms is a split in the input data, so partitions in spark are basically smaller logical chunks or divisions of the input data.

Spark distributes this partitioned data among the different nodes to perform distributed processing on the data.

10. what is spark core?

A.) spark provides distributed task scheduling, and basic I/O functionalities.

Spark uses a specialized fundamental data structure known as RDD (Resilient Distributed Datasets) that is a logical collection of data partitioned across machines.

11. what is spark driver or driver program?

A.) A Spark driver is the process that creates and owns an instance of SparkContext.

It is the cockpit of jobs and tasks execution (using DAGScheduler and Task Scheduler).

12. What is spark executors?

A.) Executors are worker nodes' that running individual tasks in a given Spark job.

They are launched at the beginning of a Spark application and typically run for the entire lifetime of an application.

13. what is worker node?

A.) Worker Node is the Slave Node. Master node assign work and worker node actually perform the assigned tasks.

Worker node processes the data stored on the node.

14. what is lazy evaluation in spark?

A.) As the name itself indicates its definition, lazy evaluation in Spark means that the execution will not start until an action is triggered.

In Spark, the picture of lazy evaluation comes when Spark transformations occur.

15. what is pair RDD in spark?

A.) Paired RDD in Spark is an RDD with the distributed collection of objects containing key-value pairs.

Paired RDDs is a very powerful data structure because it supports to act on each key operation in parallel or re-group data across the network.

16. Difference between persisit() and cache() in spark?

A.) Both persist () and cache () are the Spark optimization technique, used to store the data, but only difference is cache () method by default stores the data in-memory (MEMORY_ONLY) whereas in persist () method developer can define the storage level to in-memory or in-disk.

17. what is serialization and deserialization?

A.) serialization can be defined as which is converting from object to bytes which can be sent over network.

deserialization can be defined as converting the data to a form that can be stored sent over the network into a form which can be read.

18. Avoid returning null, in scala code,

usin None which in turn can very well handled by getOrElse

19. Diamond problem in scala occurs when child class/object tries to refer?

A.) multiple parent classes having same method name.

20. Singleton, Lazy initialization design patterns to for respectively if we have

memory constraints, defer expensive computation.

21. For the following code in scala: lazy val output = {println(“Hello”); 1} println(“Learning Scala”) println(output). What can be the result, in proper order?

A.) Learning scala, Hello,1

22. Suppose we have a series of 9 Mapreduce Jobs, then how many Disk I/Os are needed in total?

A.) 18

23. Which operations is not lazy?

A.) collect, take

24. Suppose while running spark on hadoop2, input file is of size 800 MB. How many RDD partitions will be created in all ?

A.) 7 partitions

25. which will help Rdds to achieve resiliency?

A.) RDDs maintain a Lineage graph

RDD contents cannot be changed

26. RDDs says materialized in which condititon?

A.) when action is called to execute file with collect.

27. flatmap does not provide always multiple inputs to get multiple outputs

reduceByKey is not an action

reduceByKey cannot take two or more parameters.

we cannot create spark-context in spark-shell

28. Actions are functions applied on RDD, resulting into another RDD.

A.) true

29. Spark transformations & actions are evaluated lazily?

A.) False

30. what is higher order functions?

A.) map(), reduce(), foreach()

31. Which of the below gives one to one mapping between input & output. *?

A.) map

32. by default spark UI is available on which port?

A.) port 4040

33. what is broadcast variable?

A.) Broadcast variables in Apache Spark is a mechanism for sharing variables across executors that are meant to be read-only. Without broadcast variables these variables would be shipped to each executor for every transformation and action, and this can cause network overhead.

34. what is accumulator?

A.) it is a shared variable a single file kept in driver program and remaning executor update it.

None of the executors can read the value of accumulator, but it can only update it.

these are similar to counters in mapRduce.

19. Difference between map() and flatmap()?

A.) Map () operation applies to each element of RDD and it returns the result as new RDD. In the Map, operation developer can define his own custom business logic. While FlatMap () is similar to Map, but FlatMap allows returning 0, 1 or more elements from map function.

20. what are the various level of persistance in spark?

A.) Spark has various persistence levels to store the RDDs on disk or in memory or as a combination of both with different replication levels namely: MEMORY_ONLY; MEMORY_ONLY_SER; MEMORY_AND_DISK; MEMORY_AND_DISK_SER, DISK_ONLY; OFF_HEAP

21. What is accumulator in spark?

A.) sparkContext.accumulator () is used to define accumulator variables.

value property on the accumulator variable is used to retrieve the value from the accumulator.

Accumulators are variables that are used for aggregating information across the executors.

22. what is broadcast variable in spark?

A.) Broadcast variables in Apache Spark is a mechanism for sharing variables across executors that are meant to be read-only. Without

broadcast variables these variables would be shipped to each executor for every transformation and action, and this can cause network overhead.

23. what is checkpointing in spark?

A.) Checkpointing stores the rdd physically to hdfs and destroys the lineage that created it.

The checkpoint file won't be deleted even after the Spark application terminated.

Checkpoint files can be used in subsequent job run or driver program

24. what is spark context?

A.) SparkContext is an entry point to Spark and defined in org.apache.spark package and used to programmatically create Spark RDD, accumulators, and broadcast variables on the cluster. Its object sc is default variable available in spark-shell and it can be programmatically created using SparkContext class.

25. what is Executor memory in spark?

A.) The heap size is what referred to as the Spark executor memory which is controlled with the spark.executor.memory property of the –executor-memory flag. Every spark application will have one executor on each worker node. The executor memory is basically a measure on how much memory of the worker node will the application utilize

26. Explain spark stages?

A.) Spark stages are the physical unit of execution for the computation of multiple tasks. The Spark stages are controlled by the Directed Acyclic Graph (DAG) for any data processing and transformations on the resilient distributed datasets (RDD).

Basically, there are two types of stages in spark- ShuffleMapstage and ResultStage.

27. what is spark SQL?

A.) Spark SQL is a Spark module for structured data processing. It provides a programming abstraction called DataFrames and can also act as a distributed SQL query engine. It enables unmodified Hadoop Hive queries to run up to 100x faster on existing deployments and data.

28. Difference between RDD vs Dataframe & Dataset in spark?

A.) The schema gives an expressive way to navigate inside the data. RDD is a low level API whereas DataFrame/Dataset are high level APIs. With RDD, you have more control on what you do. A DataFrame/Dataset tends to be more efficient than an RDD. What happens inside Spark core is that a DataFrame/Dataset is converted into an optimized RDD.

29. How spark SQL is different from HQL & SQL?

A.) Spark-sql: SparkSQL is a special component on the sparkCore engine that support SQL and HiveQueryLanguage without changing any syntax

SQL - SQL is a traditional query language that directly interacts with RDBMs

HQL - HQL is a JAVA-based OOP language that uses the Hibernate interface to convert the OOP code into query statements and then interacts with databases.

30. What is catalyst Optimizer?

A.) Catalyst optimizer makes use of some advanced programming language features to build optimized queries. Catalyst optimizer was developed using programming construct Scala. Catalyst Optimizer allows both Rule Based optimization and Cost Based optimization.

31. what is spark streaming?

A.) Spark Streaming is an extension of the core Spark API that allows data engineers and data scientists to process real-time data from various sources including (but not limited to) Kafka, Flume, and Amazon Kinesis. This processed data can be pushed out to file systems, databases.

32. What is DStream?

A.) DStream is a continuous stream of data. It receives input from various sources like Kafka, Flume, Kinesis, or TCP sockets. It can also

be a data stream generated by transforming the input stream. At its core, DStream is a continuous stream of RDD (Spark abstraction).

33. How to create Micro batch and its benifit?

A.) It allows developers to write code, and influence the architecture.

Microservices are small applications that your development teams create independently.

34. what is windowing in spark streaming?

A.) Window The simplest windowing function is a window, which lets you create a new DStream, computed by applying the windowing parameters to the old DStream.

35. what is Scala programming Languages & its advantages?

A.) Easy to Pick Up

Pretty Good IDE support

IntelliJ IDEA. Most developers consider this to be the best IDE for Scala. It has great UI, and the editor is pretty...

Scalability

36. What is the difference between Statically typed & Dynamically typed language?

A.) In statically typed languages type checking happens at compile time. Whereas, in dynamically typed languages, type checking happens at run-time.

37. what is the difference var and val in scala?

A.) The keywords var and val both are used to assign memory to variables.

var keyword initializes variables that are mutable, and the val keyword initializes variables that are immutable.

38. what is the difference between == in java and scala?

A.) In Java, C++, and C# the == operator tests for reference, not value equality.

in Scala, == is testing for value equality.

39. What is Typesafe in scala?

A.) Type-safe means that the set of values that may be assigned to a program variable must fit well-defined and testable criteria.

40. what is type inference in scala?

A.) With Scala type inference, Scala automatically detects the type of the function without explicitly specified by the user.

41. what is Unit in scala? what is difference between java void's and scala unit?

A.) Unit is a final class defined in “scala” package that is “scala.Unit”. Unit is something similar to Java’s void. But they have few differences. Java’s void does not any value. It is nothing. Scala’s Unit has one value () is the one and only value of type Unit in Scala.

42. what is scala singleton object?

A.) Scala Singleton Object Singleton object is an object which is declared by using object keyword instead by class. No object is required to call methods declared inside singleton object. In scala, there is no static concept.

43. what is companion object in scala?

A.) A companion object in Scala is an object that’s declared in the same file as a class, and has the same name as the class. For instance, when the following code is saved in a file named Pizza.scala, the Pizza object is considered to be a companion object to the Pizza class:

44. Difference between and singleton object and class in scala?

A.) An object is a singleton -- an instance of a class which is guaranteed to be unique. For every object in the code, an anonymous class is created, which inherits from whatever classes you declared object to implement.

Classes have fields and methods

45. what is scala Map?

A.) Scala map is a collection of key/value pairs. Any value can be retrieved based on its key. Keys are unique in the Map, but values need not be unique.

46. what is scala set?

A.) Set is a collection that contains no duplicate elements. There are two kinds of Sets, the immutable and the mutable.

47. what is the use of Tuples in scala?

A.) A tuple is a data structure which can store elements of the different data type. It is also used for storing and retrieving of data. In scala, tuples are immutable in nature and store heterogeneous types of data.

48. what is Scala case class?

A.) Scala Case Class is like a regular class, except it is good for modeling immutable data. It also serves useful in pattern matching, such a class has a default apply () method which handles object construction.

49. what is scala option?

A.) Scala Option [T] is a container for zero or one element of a given type. An Option [T] can be either Some [T] or None object, which represents a missing value.

50. what is use case of App class in scala?

A.) Scala provides a helper class, called App, that provides the main method. Instead of writing your own main method, classes can extend the App class to produce concise and executable applications in Scala.

51. Difference between terms & types in scala? Nill, NUll, None, Nothing?

A.) Null– Its a Trait. null– Its an instance of Null- Similar to Java null.

Nil– Represents an emptry List of anything of zero length.

Nothing is a Trait. Its a subtype of everything. But not superclass of anything. There are no instances of Nothing.

None– Used to represent a sensible return value. Just to avoid null pointer.

52. what is traits in scala?

A.) In scala, trait is a collection of abstract and non-abstract methods. You can create trait that can have all abstract methods or some abstract and some non-abstract methods. A variable that is declared either by using val or var keyword in a trait get internally implemented in the class that implements the trait.

53. Difference between Traits and abstract class in scala?

A.) Traits Abstract Class

Allow multiple inheritances. Do not Allow multiple inheritances.

Constructor parameters are not allowed in Trait. Constructor parameter are allowed in Abstract Class.

The Code of traits is interoperable until it is implemented. The code of abstract class is fully interoperable.

Traits can be added to an object instance in Scala. Abstract classes cannot be added to object instance in Scala.

54. Difference between Call-by-value and call-by-name parameter?

A.) call-by-value is The same value will be used all throughout the function. Whereas in a Call by Name, the expression itself is passed as a parameter to the function and it is only computed inside the function, whenever that particular parameter is called.

55. what are Higher order functions in scala?

A.) Scala Higher Order Functions Higher order function is a function that either takes a function as argument or returns a function. In other words we can say a function which works with function is called higher order function. Higher order function allows you to create function composition, lambda function or anonymous function etc.

56. What is Pure function in scala?

A.) A function is called pure function if it always returns the same result for same argument values and it has no side effects like modifying an argument (or global variable) or outputting something.

57. Explain scala anonymous function in scala?

A.) In Scala, An anonymous function is also known as a function literal. A function which does not contain a name is known as an anonymous function. An anonymous function provides a lightweight function definition.

58. what is closure in scala?

A.) A closure is a function, whose return value depends on the value of one or more variables declared outside this function.

59. what is currying in scala?

A.) Currying is the process of converting a function with multiple arguments into a sequence of functions that take one argument.

60. what is option in scala? why do we use it?

A.) Scala Option[T] is a container for zero or one element of a given type. An Option[T] can be either Some[T] or None object, which represents a missing value.

Option type is used frequently in Scala programs and you can compare this with the null value

61. what is tail recursion in scala?

A.) Recursion is a method which breaks the problem into smaller subproblems and calls itself for each of the problems.

62. What is yield in scala?

A.) For each iteration of your for loop, yield generates a value which is remembered by the for loop (behind the scenes)

63. can we able to do datasets in python?

A.) A simple way to get sample datasets in Python is to use the pandas ‘read_csv’ method to load them directly from the internet.

64. How to join two tables using dataframes?

A.) empDF. join (deptDF, empDF ("emp_dept_id") === deptDF ("dept_id"),"inner").show(false)

65. How to remove duplicates records in dataframe?

A.) Use distinct () – Remove Duplicate Rows on DataFrame.

Use dropDuplicate () – Remove Duplicate Rows on DataFrame.

66. How to add columns in Dataframe?

A.) Using withColumn () to Add a New Column. Here, we have added a new column.

67. SQL basics concepts such as Rank, Dense Rank, Row Number?

A.) RANK() Returns the rank of each row in the result set of partitioned column. select Name,Subject,Marks, RANK()

DENSE_RANK() This is same as RANK() function. Only difference is returns rank without gaps.

ROW_NUMBER will always generate unique values without any gaps, even if there are ties.

68. Query to find 2nd largest number in the table?

A.) SELECT MAX(sal) as Second_Largest FROM emp_test WHERE sal < (SELECT MAX(sal) FROM emp_test)

69. To find duplicate record in table?

A.) select a.* from Employee a where rowid != (select max(rowid) from Employee b where a.Employee_num =b.Employee_num;

70. Difference between list and Tuple?

A.) LIST	TUPLE
Lists are mutable	Tuples are immutable

Implication of iterations is Time-consuming The implication of iterations is comparatively Faster

The list is better for performing operations, such as insertion and deletion. Tuple data type is appropriate for accessing the elements

Lists consume more memory Tuple consume less memory as compared to the list

Lists have several built-in methods Tuple does not have many built-in methods.

The unexpected changes and errors are more likely to occur In tuple, it is hard to take place.

71. Difference between def and Lambda?

A.) lambda is a keyword that returns a function object and does not create a 'name'. Whereas def creates name in the local namespace

lambda functions are good for situations where you want to minimize lines of code as you can create function in one line of python code. ...

lambda functions are somewhat less readable for most Python users.

72. Why bigdata on cloud preferred these days?

A.) Big Data Cloud brings the best of open source software to an easy-to-use and secure environment that is seamlessly integrated and serverless.

73. What is aws EMR?

A.) Amazon Elastic MapReduce (EMR) is an Amazon Web Services (AWS) tool for big data processing and analysis. Amazon EMR offers the expandable low-configuration service as an easier alternative to running in-house cluster computing.

74. How to write a UDF in hive?

A.) By writing UDF (User Defined function) hive makes it easy to plug in your own processing code and invoke it from a Hive query. UDF's have to be writhen in Java, the Language that Hive itself is written in.

Create a Java class for the User Defined Function which extends ora.apache.hadoop.hive.sq.exec.UDF and implements more than one evaluate () methods. Put in your desired logic and you are almost there.

Package your Java class into a JAR file (I am using Maven)

Go to Hive CLI, add your JAR, and verify your JARs is in the Hive CLI classpath

CREATE TEMPORARY FUNCTION in Hive which points to your Java class

Use it in Hive SQL

75. file formats row based vs column based?

A.) In a row storage format, each record in the dataset has to be loaded, parsed into fields and then the data for Name is extracted. With the column-oriented format, it can directly go to the Name

column as all the values for that column are stored together. It doesn't need to go through the whole record.

76. What is RDD?

A.) Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark. It is an immutable distributed collection of objects. Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster.

77. how to join two larger tables in spark?

A.) Spark uses SortMerge joins to join large table. It consists of hashing each row on both table and shuffle the rows with the same hash into the same partition. There the keys are sorted on both side and the sortMerge algorithm is applied.

78. what is Bucketed tables?

A.) In the table directory, the Bucket numbering is 1-based and every bucket is a file. Bucketing is a standalone function. This means you can perform bucketing without performing partitioning on a table. A bucketed table creates nearly equally distributed data file sections.

79. How to read the parquet file format in spark?

A.) Similar to write, DataFrameReader provides parquet () function (spark.read.parquet) to read the parquet files and creates a Spark DataFrame.

80. How to tune spark executor, cores and executor memory?

A.) Number of cores = Concurrent tasks as executor can run

total available cores in one Node(CPU) - we come to

3 executors per node.

The property spark.executor.memory specifies the amount of memory to allot to each executor.

81. Default partition size in spark?

A.) default number of partitions is based on the following. On the HDFS cluster, by default, Spark creates one Partition for each block of the file. In Version 1 Hadoop the HDFS block size is 64 MB and in Version 2 Hadoop the HDFS block size is 128 MB

82. is there any use of running spark program on single machine?

A.) Spark also provides a simple standalone deploy mode. You can launch a standalone cluster either manually, by starting a master and workers by hand, or use our provided launch scripts. It is also possible to run these daemons on a single machine for testing.

83. how to find how many resources are available in YARN?

A.) yarn.resource-types.memory-mb.increment-allocation: The fairscheduler grants memory in increments of this value. If you submit a task with resource request that is not a multiple of

memory-mb.increment-allocation, the request will be rounded up to the nearest increment. Defaults to 1024 MB. yarn.resource-types.vcores.increment-allocation

84. Differences between cluster and client Mode?

A.) In cluster mode, the driver will get started within the cluster in any of the worker machines. So, the client can fire the job and forget it. In client mode, the driver will get started within the client. So, the client has to be online and in touch with the cluster.

85. Explain about the dynamic allocation in spark?

A.) Spark dynamic allocation is a feature allowing your Spark application to automatically scale up and down the number of executors. And only the number of executors not the memory size and not the number of cores of each executor that must still be set specifically in your application or when executing spark-submit command.

86. Difference between partition by and cluster by in hive?

A.) In Hive partitioning, the table is divided into the number of partitions, and these partitions can be further subdivided into more manageable parts known as Buckets/Clusters. Records with the same bucketed column will be stored in the same bucket. “clustered by” clause is used to divide the table into buckets.

Cluster By is a short-cut for both Distribute By and Sort By. Hive uses the columns in Distribute By to distribute the rows among reducers.

All rows with the same Distribute By columns will go to the same reducer. However, Distribute By does not guarantee clustering or sorting properties on the distributed keys.

87. How to choose partitioning column in hive? and which column shouldn't use partition and why?

A.) When the column with a high search query has low cardinality. For example, if you create a partition by the country name then a maximum of 195 partitions will be made and these number of directories are manageable by the hive. On the other hand, do not create partitions on the columns with very high cardinality.

88. how to transfer data from unix system to HDFS?

A.) hdfs dfs -put test /hadoop ubuntu@ubuntu-VirtualBox

89. can we extract only different data from two different tables?

A.) using Join column we can extract data.

SELECT tablenmae1.colunmname, tablename2.columnnmae

FROM tablenmae1

JOIN tablename2

ON tablenmae1.colunmnam = tablename2.columnnmae

ORDER BY columnname;

90. What is the difference between SQL vs NoSQL?

A.) SQL databases are vertically scalable while NoSQL databases are horizontally scalable. SQL databases have a predefined schema whereas NoSQL databases use dynamic schema for unstructured data. SQL requires specialized DB hardware for better performance while NoSQL uses commodity hardware.

91. how to find particular text name in HDFS?

A.) You can use cat command on HDFS to read regular text files. hdfs dfs -cat /path/to/file.csv

92. Explain about sqoop ingestion process?

A.) Apache Sqoop is a data ingestion tool designed for efficiently transferring bulk data between Apache Hadoop and structured data-stores such as relational databases, and vice-versa.

93. Explain about sort Merge Bucket Join?

A.) Sort Merge Bucket (SMB) join in hive is mainly used as there is no limit on file or partition or table join. SMB join can best be used when the tables are large. In SMB join the columns are bucketed and sorted using the join columns. All tables should have the same number of buckets in SMB join.

94. Explain about tungsten?

A.) Tungsten is a Spark SQL component that provides increased performance by rewriting Spark operations in bytecode, at runtime.

95. How can we join two bigger tables in spark?

A.) either using Sort Merge Joins if we are joining two big tables, or Broadcast Joins if at least one of the datasets involved is small enough to be stored in the memory of the single all executors.

A ShuffleHashJoin is the most basic way to join tables in Spark

96. Explain about left outer join?

A.) The left outer join returns a resultset table with the matched data from the two tables and then the remaining rows of the left table and null from the right table's columns.

97. How to count the lines in a file by using linux command?

A. using -wc

98. How to achieve map side joins in hive?

A.) only possible since the right table that is to the right side of the join conditions, is lesser than 25 MB in size. Also, we can convert a right-outer join to a map-side join in the Hive.

99. when we use select command does it goes to reducer in Hive?

A.) We can use reducer if and there is no aggregation of data in mapside only then it uses reducer.

100. How to validate the data once the ingestion is done?

A.) data validation is used as a part of processes such as ETL (Extract, Transform, and Load) where you move data from a source database to a target data warehouse so that you can join it with other data for analysis. Data validation helps ensure that when you perform analysis, your results are accurate.

Steps to data validation:

a. Determine data sample: validate a sample of your data rather than the entire set.

b. Validata the database: Before you move your data, you need to ensure that all the required data is present in your existing database

c. Validate the data format: Determine the overall health of the data and the changes that will be required of the source data to match the schema in the target.

Methods for data validation:

a. Scripting: Data validation is commonly performed using a scripting language

b. Enterprise tools: Enterprise tools are available to perform data validation.

c. open source tools: Open source options are cost-effective, and if they are cloud-based, can also save you money on infrastructure costs.

101. what is the use of split by command in sqoop?

A.) split-by in sqoop is used to create input splits for the mapper. It is very useful for parallelism factor as splitting imposes the job to run faster. Hadoop MAP Reduce is all about divide and conquer. When using the split-by option, you should choose a column which contains values that are uniformly distributed.

102. Difference between dataframe vs datasets?

A.) DataFrames gives a schema view of data basically, it is an abstraction. In dataframes, view of data is organized as columns with column name and types info. In addition, we can say data in dataframe is as same as the table in relational database.

As similar as RDD, execution in dataframe too is lazy triggered.

In Spark, datasets are an extension of dataframes. Basically, it earns two different APIs characteristics, such as strongly typed and untyped. Datasets are by default a collection of strongly typed JVM objects, unlike dataframes. Moreover, it uses Spark's Catalyst optimizer.

103. Difference between schema on read vs schema on write?

A.) Schema on read differs from schema on write because you create the schema only when reading the data. Structured is applied to the data only when it's read, this allows unstructured data to be stored in the database.

The main advantages of schema on write are precision and query speed.

104. Different types of partition in hive?

A.) Types of Hive Partitioning

Static Partitioning

Dynamic Partitioning

105. How to find counts based on age group?

A.) SELECT Col1, COUNT(*) FROM Table GROUP BY Col1.

106. How to find a word in a log file by using pyspark?

A.) input_file = sc.textFile("/path/to/text/file")

map = input_file.flatMap(lambda line: line.split(" ")).map(lambda word: (word, 1))

counts = map.reduceByKey(lambda a, b: a + b)

counts.saveAsTextFile("/path/to/output/")

107. Explain about Executor node in spark?

A.) Executors are worker nodes' processes in charge of running individual tasks in a given Spark job. They are launched at the beginning of a Spark application and typically run for the entire lifetime of an application. Once they have run the task they send the results to the driver.

108. Difference between Hadoop & spark?

A.) biggest difference is that it works in memory. Whereas Hadoop reads and writes files to HDFS, Spark processes data in RAM using a concept known as an RDD, Resilient Distributed Dataset.

Hadoop has its own storage system HDFS while Spark requires a storage system like HDFS which can be easily grown by adding more nodes. They both are highly scalable as HDFS storage can go more than hundreds of thousands of nodes. Spark can also integrate with other storage systems like S3 bucket.

109. Query to find duplicate value in SQL?

A.) Using the GROUP BY clause to find the duplicate values

use the GROUP BY clause to group all rows by the target column, which is the column that you want to check duplicate. Then, use the COUNT ()

110. Difference between Row number and Dense Rank in SQL?

A.) Rank () SQL function generates rank of the data within ordered set of values but next rank after previous rank is row_number of that particular row. On the other hand, Dense_Rank () SQL function generates next number instead of generating row_number. Below is the SQL example which will clarify the concept

111. What are the various hive optimization techniques?

A.) Tez-Execution Engine in Hive:Tez Execution Engine – Hive Optimization Techniques, to increase the Hive performance.

Usage of Suitable File Format in Hive

Hive Partitioning.

Bucketing in Hive.

112. How Mapreduce will work? Explain?

A.) MapReduce can perform distributed and parallel computations using large datasets across a large number of nodes. A MapReduce job usually splits the input datasets and then process each of them independently by the Map tasks in a completely parallel manner. The output is then sorted and input to reduce tasks.

it uses key value pair.

113. How many types of tables have in Hive?

A.) Hive knows two different types of tables: Internal table and the External table. The Internal table is also known as the managed table.

114. How to drop table in HBase?

A.) Dropping a Table using HBase Shell Using the drop command, you can delete a table. Before dropping a table, you have to disable it. hbase (main):018:0> disable 'emp' 0 row (s) in 1.4580 seconds hbase (main):019:0> drop 'emp' 0 row (s) in 0.3060 seconds

115. What is the difference between Batch and real time processing?

A.) In real time processing processor needs to very responsive and active all the time.

In batch processing processor only needs to busy when work is assigned to it.

Real-time processing needs high computer architecture and high hardware specification.

Normal computer specification can also work with batch processing.

Time to complete the task is very critical in real-time.

Real-time processing is expensive.

Batch processing is cost effective

116. How can Apache spark be used alongside Hadoop?

A.) Running Alongside Hadoop You can run Spark alongside your existing Hadoop cluster by just launching it as a separate service on the same machines. To access Hadoop data from Spark, just use a hdfs:// URL (typically hdfs://<namenode>:9000/path, but you can find the right URL on your Hadoop Namenode's web UI).

117. What is Data explode and lateral view in Hive?

A.) In Hive, lateral view explode the array data into multiple rows. In other word, lateral view expands the array into rows.

Hive has way to parse array data type using LATERAL VIEW. Use LATERAL VIEW with UDTF to generate zero or more output rows for each input row. Explode is one type of User Defined Table Building Function.

118. What is combiner, shuffling, sorting in Mapreduce?

A.) Shuffling is the process by which it transfers mappers intermediate output to the reducer. Reducer gets 1 or more keys and associated values on the basis of reducers.

In Sort phase merging and sorting of map output takes place.

The combiner should combine key/value pairs with the same key. Each combiner may run zero, once, or multiple times.

119. what is the difference between reduceByKey and GroupByKey?

A.) The groupByKey method operates on an RDD of key-value pairs, so key a key generator function is not required as input. What is reduceByKey? The reduceByKey is a higher-order method that takes associative binary operator as input and reduces values with the same key. This function merges the values of each key using the reduceByKey method in Spark.

120. what is static and dynamic partition in Hive?

A.) Usually dynamic partition load the data from non partitioned table. Dynamic Partition takes more time in loading data compared to static partition. When you have large data stored in a table then Dynamic partition is suitable.

Static partitions are preferred when loading big data in Hive table and it saves your time in loading data compared to dynamic partition. In static partitioning, we need to specify the partition column value in each and every LOAD statement.

121. Udf example in Hive?

A.) A UDF processes one or several columns of one row and outputs one value. For example : SELECT lower(str) from table For each row in "table," the "lower" UDF takes one argument, the value of "str", and outputs one value, the lowercase representation of "str".

122. what is Serde in Hive?

A.) In SerDe interface handles both serialization and deserialization and also interpreting the results of serialization as individual fields for processing. It allows Hive to read data from a table, and write it back to HDFS in any format, user can write data formats.

123. How to check the file size in Hadoop?

A.) You can use the hadoop fs -ls -h command to check the size. The size will be displayed in bytes.

124. How to submit the spark Job?

A.) Using --deploy-mode, you specify where to run the Spark application driver program.

cluster Mode: In cluster mode, the driver runs on one of the worker nodes, and this node shows as a driver on the Spark Web UI of your application. cluster mode is used to run production jobs.

Client mode: In client mode, the driver runs locally where you are submitting your application from. client mode is majorly used for interactive and debugging purposes.

Using --master option, you specify what cluster manager to use to run your application. Spark currently supports Yarn, Mesos, Kubernetes, Stand-alone, and local.

125. what is vectorization and why it used?

A.) Vectorization is the process of converting an algorithm from operating on a single value at a time to operating on a set of values at one time

126. what are Complex data types in Hive?

A.) ARRAY

Struct

Map

127. What is sampling in Hive?

A.) Prepare the dataset.

Create a Hive Table and Load the Data

Sampling using Random function.

Create a Bucketed table

Load data into Bucketed table

Sampling using bucketing.

Block sampling in hive.

128. what are different type of xml files in Hadoop?

A.) resource-types.xml

node-resources.xml

yarn-site.xml

129. what is case class?

A.) A Case Class is just like a regular class, which has a feature for modeling unchangeable data. It is also constructive in pattern matching.

It has been defined with a modifier case , due to this case keyword, we can get some benefits to stop oneself from doing a sections of codes that have to be included in many places with little or no alteration.

130. Difference between broadcast and accumulators?

A.) Broadcast variables − used to efficiently, distribute large values. Accumulators − used to aggregate the information of particular collection.

131. what is the difference between spark context and spark session?

A.) SparkContext has been available since Spark 1.x versions and it's an entry point to Spark when you wanted to program and use Spark RDD. Most of the operations/methods or functions we use in Spark are comes from SparkContext for example accumulators, broadcast variables, parallelize and more.

SparkSession is essentially combination of SQLContext, HiveContext and future StreamingContext.

132. what are the operation of dataframe?

A.) A spark data frame can be said to be a distributed data collection that is organized into named columns and is also used to provide the operations such as filtering, computation of aggregations, grouping and also can be used with Spark SQL.

133. Explain why spark preferred over mapreduce?

A.) the benefits of Apache Spark over Hadoop MapReduce are given below: Processing at high speeds: The process of Spark execution can be up to 100 times faster due to its inherent ability to exploit the memory rather than using the disk storage.

134. what is the difference between partitioning and Bucketing?

A.) Bucketing decomposes data into more manageable or equal parts. With partitioning, there is a possibility that you can create

multiple small partitions based on column values. If you go for bucketing, you are restricting number of buckets to store the data.

135. How to handle incremental data in bigdata?

A.) Move existing HDFS data to temporary folder

Run last modified mode fresh import

Merge with this fresh import with old data which saved in temporary folder.

136. Explain the Yarn Architecture?

A.) Apache YARN framework contains a Resource Manager (master daemon), Node Manager (slave daemon), and an Application Master.

YARN is the main component of Hadoop v2.0. YARN helps to open up Hadoop by allowing to process and run data for batch processing, stream processing, interactive processing and graph processing which are stored in HDFS. In this way, It helps to run different types of distributed applications other than MapReduce.

137. what is incremental sqoop?

A.) Incremental imports mode can be used to retrieve only rows newer than some previously-imported set of rows. Why Append mode ?? works for numerical data that is incrementing over time, such as auto-increment keys,

138. Why we use Hbase and how it store data?

A.) HBase provides a flexible data model and low latency access to small amounts of data stored in large data sets HBase on top of Hadoop will increase the throughput and performance of distributed cluster set up. In turn, it provides faster random reads and writes operations

139. What are various optimization technique in hive?

A.) Apache Hive Optimization Techniques — 1. Partitioning. Bucketing.

140. Sqoop command to exclude tables while retrieval?

A.) sqoop import --connect jdbc:mysql://localhost/sqoop --username root --password hadoop --table <tablename> --target-dir '/Sqoop21/AllTables' --exclude-tables <table1>,<tables2>.

141. how to create sqoop password alias?

A.) sqoop import --connect jdbc:mysql://database.example.com/employees \ --username dbuser --password-alias mydb.password.alias. Similarly, if the command line option is not preferred, the alias can be saved in the file provided with --password-file option.

142. what is sqoop job optimization?

A.) To optimize performance, set the number of map tasks to a value lower than the maximum number of connections that the database supports.

143. what is sqoop boundary quieries and split by usage?

A.) The boundary query is used for splitting the value according to id_no of the database table. To boundary query, we can take a minimum value and maximum value to split the value.

split-by in sqoop is used to create input splits for the mapper. It is very useful for parallelism factor as splitting imposes the job to run faster.

144. what is hbase compaction technique and wirte operation hbase using spark??

A.) HBase Minor Compaction The procedure of combining the configurable number of smaller HFiles into one large HFile is what we call Minor compaction.

hbase-spark connector which provides HBaseContext to interact Spark with HBase. HBaseContext pushes the configuration to the Spark executors and allows it to have an HBase Connection per Spark Executor.

145. what are hive managed Hbase tables and how to create that?

A.) Hive tables Managed tables are Hive owned tables where the entire lifecycle of the tables' data are managed and controlled by Hive. External tables are tables where Hive has loose coupling with

the data. Replication Manager replicates external tables successfully to a target cluster.

CREATE [EXTERNAL] TABLE foo(…) STORED BY
‘org.apache.hadoop.hive.hbase.HBaseStorageHandler’
TBLPROPERTIES (‘hbase.table.name’ = ‘bar’); [/sql]

146. How Hbase can be a Distributed database?

A.) Hbase is one of NoSql column-oriented distributed database available in apache foundation. HBase gives more performance for retrieving fewer records rather than Hadoop or Hive. It's very easy to search for given any input value because it supports indexing, transactions, and updating.

147. What is hive metastore and how to access that?

A.) Metastore is the central repository of Apache Hive metadata. It stores metadata for Hive tables (like their schema and location) andpartitions in a relational database. It provides client access to this information by using metastore service API. Hive metastore consists of two fundamental units: 1. A service th

at provides metastore access to other Apache Hive services. 2. Disk storage for the Hive metadata which is separate from HDFSstorage.

probably get most of the information you need through HCatalog, without direct access to the metastore tables.

148. What are the ways to remove duplicates in hive?

A.) Use Insert Overwrite and DISTINCT Keyword

GROUP BY Clause to Remove Duplicate

Use Insert Overwrite with row_number () analytics functions

148. What is Hive Managed and External tables?

A.) Managed tables are Hive owned tables where the entire lifecycle of the tables' data are managed and controlled by Hive. External tables are tables where Hive has loose coupling with the data. Replication Manager replicates external tables successfully to a target cluster. The managed tables are converted to external tables after replication.

149. How partition can be restored?

A.) using MSCK REPAIR

150. what is data loading in hive?

A.) Hive provides us the functionality to load pre-created table entities either from our local file system or from HDFS. The LOAD DATA statement is used to load data into the hive table.

151. How to automate Hive jobs?

A.) Like you can also use Hive CLI and its very ease to do such jobs. You can write shell script in Linux or .bat in Windows. In script you can simply go like below entries. $HIVE_HOME/bin/hive -e 'select

a.col from tab1 a'; or if you have file : $HIVE_HOME/bin/hive -f /home/my/hive-script.sql Make sure you have set $HIVE_HOME in your env.

152. where do we run job in spark?

A.) The spark-submit script in Spark's bin directory is used to launch applications on a cluster.

153. How to allocate resources in spark?

A.) Resources allocation - Dynamic/Static; Upstream or Downstream application .

154. Difference between Edge node vs Data Node?

A.) The majority of work is assigned to worker nodes. Worker node store most of the data and perform most of the calculations Edge nodes facilitate communications from end users to master and worker nodes.

155. what is Hive context?

A.) HiveContext is a super set of the SQLContext. Additional features include the ability to write queries using the more complete HiveQL parser, access to Hive UDFs, and the ability to read data from Hive tables. And if you want to work with Hive you have to use HiveContext, obviously.

156. How to read file from hdfs or other sources in spark?

A.) Use textFile() and wholeTextFiles() method of the SparkContext to read files from any file system and to read from HDFS, you need to provide the hdfs path as an argument to the function.

157. How to add custom schema to rdd?

A.) In spark, schema is array StructField of type StructType. Each StructType has 4 parameters. Column Name ; Data type of that column; Boolean value indication if values in this column can be null or not; Metadata column - this is optional column which can be used to add additional information about column

158. How to convert dataframe to rdd?

A.) To convert a dataframe back to rdd simply use the .rdd method: rdd = df.rdd But the setback here is that it may not give the regular spark RDD, it may return a Row object. In order to have the regular RDD format run the code below: rdd = df.rdd.map(tuple)

159. Difference between case class and class?

A.) A class can extend another class, whereas a case class can not extend another case class (because it would not be possible to correctly implement their equality).

160. what is optimization technique in spark?

A.) Spark optimization techniques are used to modify the settings and properties of Spark to ensure that the resources are utilized properly and the jobs are executed quickly. All this ultimately helps in processing data efficiently. The most popular Spark optimization techniques are listed below:

popular Spark optimization techniques are listed:

1. Data Serialization: Here, an in-memory object is converted into another format that can be stored in a file or sent over a network.

a. Java serialization: The ObjectOutputStream framework is used for serializing objects.

b. kyro serialization : To improve the performance, the classes have to be registered using the registerKryoClasses method.

2. caching: This is an efficient technique that is used when the data is required more often. Cache() and persist() are the methods used in this technique.

3. Data structure tuning: We can reduce the memory consumption while using Spark, by tweaking certain Java features that might add overhead.

4. Garbage collection optimization: G1 and GC must be used for running Spark applications. The G1 collector manages growing heaps. GC tuning is essential according to the generated logs, to control the unexpected behavior of applications.

161. What is unit data type in scala?

A.) The Unit type in Scala is used as a return statement for a function when no value is to be returned. Unit type can be e compared to void data type of other programming languages like Java.

162. What is boundary query in sqoop?

A.) The boundary query is used for splitting the value according to id_no of the database table. To boundary query, we can take a minimum value and maximum value to split the value. To make split using boundary queries, we need to know all the values in the table.

163. What is the use of sqoop eval command?

A.) It allows users to execute user-defined queries against respective database servers and preview the result in the console. So, the user can expect the resultant table data to import. Using eval, we can evaluate any type of SQL query that can be either DDL or DML statement.

164. How can we decide number of bucketing?

A.) The number of buckets is determined by hashFunction (bucketingColumn) mod numOfBuckets numOfBuckets is chose when you create the table with partitioning. The hash function output depends on the type of the column choosen.

165. Is it possible to bucketing and partitioning on same column?

A.) Yes.

Partitioning is you data is divided into number of directories on HDFS. Each directory is a partition.

166. How to do optimized joins in Hive?

A.) Use Tez to Fasten the execution Apache TEZ is an execution engine used for faster query execution.

Enable compression in Hive Compression techniques reduce the amount of data being transferred

Use ORC file format ORC (optimized record columnar) is great when it comes to hive performance tuning.

167. How to optimize join of 2 big tables?

A.) use the Bucket Map Join. For that the amount of buckets in one table must be a multiple of the amount of buckets in the other table. It can be activated by executing set hive.optimize.bucketmapjoin=true; before the query. If the tables don't meet the conditions, Hive will simply perform the normal Inner Join.

If both tables have the same amount of buckets and the data is sorted by the bucket keys, Hive can perform the faster Sort-Merge Join

168. what are major issues faced in spark development?

A.) Debugging - Spark although can be written in Scala, limits your debugging technique during compile time. You would encounter

many run-time exceptions while running the Spark job. This is, many a times, because of the data. Sometimes because of the data type mismatch (there is dynamic data type inference) and sometimes data having null values and all. So there will be lot of iterations of run-time debugging.

Optimization - Optimizing a Spark code is a job to do. You need to optimize from the code side and from the resource allocation side too. A very well written code with good logic often performs badly because of badly done resource allocation.

169. what is dynamic allocation?

A.) Dynamic partitions provide us with flexibility and create partitions automatically depending on the data that we are inserting into the table.

170. What types of transformations do we perform in spark?

A.) Narrow transformation

Wide transformation

171. how to load data in hive table?

A.) Using Insert Command

table to table load

172. Difference between Map Vs Map Partition?

A.) Mappartitions is a transformation that is similar to Map. In Map, a function is applied to each and every element of an RDD and returns each and every other element of the resultant RDD. In the case of mapPartitions, instead of each element, the function is applied to each partition of RDD

mapPartitions exercises the function at the partition level

173. If we have some header information in a file how to read from it, and how to convert it to dataset or dataframe?

A.) we can add the option like header is true in while reading the file.

174. Difference between case class vs Struct type?

A.) Structs are value types and are copied on assignment. Structs are value types while classes are reference types. Structs can be instantiated without using a new operator. A struct cannot inherit from another struct or class, and it cannot be the base of a class.

175. What is sort by vs Order by in hive?

A.) Hive sort by and order by commands are used to fetch data in sorted order. The main differences between sort by and order by commands are given below. Sort by. hive> SELECT E.EMP_ID FROM Employee E SORT BY E.empid; hive> SELECT E.EMP_ID FROM Employee E SORT BY E.empid; May use multiple reducers for final output.

176. How to increase the performance of Sqoop?

A.) Controlling the amount of parallelism that Sqoop will use to transfer data is the main way to control the load on your database.

Using more mappers will lead to a higher number of concurrent data transfer tasks, which can result in faster job completion.

However, it will also increase the load on the database as Sqoop will execute more concurrent queries.

177. While sqooping some data loss. how to handle that?

A.) Some lost data is recoverable, but this process often requires the assistance of IT professionals and costs time and resources your business could be using elsewhere. In other instances, lost files and information cannot be recovered, making data loss prevention even more essential.

Reformatting can also occur during system updates and result in data loss.

178. How to update record in Hbase table?

A.) Using put command you can insert a record into the HBase table easily. Here is the HBase Create data syntax. We will be using Put command to insert data into HBase table

179. what happens when sqoop fails in between the large data import job?

A.) sqoop import - job failure between data import, due to insert collisions in some cases, or lead to duplicated data in others. Since Sqoop breaks down export process into multiple transactions, it is possible that a failed export job may result in partial data being committed to the database.

180. what are hadoop components and their services?

A.) HDFS: Hadoop Distributed File System is the backbone of Hadoop which runs on java language and stores data in Hadoop applications. They act as a command interface to interact with Hadoop. the two components of HDFS – Data node, Name Node. Name node manages file systems and operates all data nodes and maintains records of metadata updating. In case of deletion of data, they automatically record it in Edit Log. Data Node (Slave Node) requires vast storage space due to reading and writing operations.

Yarn: It’s an important component in the ecosystem and called an operating system in Hadoop which provides resource management and job scheduling task.

Hbase: It is an open-source framework storing all types of data and doesn’t support the SQL database. They run on top of HDFS and written in java language.

HBase master, Regional Server. The HBase master is responsible for load balancing in a Hadoop cluster and controls the failover. They are responsible for performing administration role. The regional server’s role would be a worker node and responsible for reading, writing data in the cache.

Sqoop: It is a tool that helps in data transfer between HDFS and MySQL and gives hand-on to import and export data

Apache spark: It is an open-source cluster computing framework for data analytics and an essential data processing engine. It is written in Scala and comes with packaged standard libraries.

Apache Flume: It is a distributed service collecting a large amount of data from the source (webserver) and moves back to its origin and transferred to HDFS. The three components are Source, sink, and channel.

MapReduce: It is responsible for data processing and acts as a core component of Hadoop. Map Reduce is a processing engine that does parallel processing in multiple systems of the same cluster.

Apache Pig: Data Manipulation of Hadoop is performed by Apache Pig and uses Pig Latin Language. It helps in the reuse of code and easy to read and write code.

Hive: It is an open-source platform for performing data warehousing concepts; it manages to query large data sets stored in HDFS. It is built on top of the Hadoop Ecosystem. the language used by Hive is Hive Query language.

Apache Drill: Apache Drill is an open-source SQL engine which process non-relational databases and File system. They are designed to support Semi-structured databases found in Cloud storage.

Zookeeper: It is an API that helps in distributed Coordination. Here a node called Znode is created by an application in the Hadoop cluster.

Oozie: Oozie is a java web application that maintains many workflows in a Hadoop cluster. Having Web service APIs controls over a job is done anywhere. It is popular for handling Multiple jobs effectively.

181. What are important configuration files in Hadoop?

A.) HADOOP-ENV.sh ->>It specifies the environment variables that affect the JDK used by Hadoop Daemon (bin/hadoop). We...

CORE-SITE.XML ->>It is one of the important configuration files which is required for runtime environment settings of...

HDFS-SITE.XML ->>It is one of the important configuration files which is required for runtime environment settings of...

MAPRED-SITE.XML ->>It is one of the important configuration files which is required for runtime environment.

182. what is rack awareness?

A.) With the rack awareness policy's we store the data in different Racks so no way to lose our data. Rack awareness helps to maximize the network bandwidth because the data blocks transfer within the Racks. It also improves the cluster performance and provides high data availability.

183. problem with having lots of small files in HDFS? and how to overcome?

A.) Problems with small files and HDFS A small file is one which is significantly smaller than the HDFS block size (default 64MB). If you're storing small files, then you probably have lots of them (otherwise you wouldn't turn to Hadoop), and the problem is that HDFS can't handle lots of files.

Hadoop Archive

Sequence files

184. Main difference between Hadoop 1 and Hadoop 2?

A.) Hadoop 1.x System is a Single Purpose System. We can use it only for MapReduce Based Applications. If we observe the components of Hadoop 1.x and 2.x, Hadoop 2.x Architecture has one extra and new component that is : YARN (Yet Another Resource Negotiator).

185. What is block scanner in hdfs?

A.) Block Scanner is basically used to identify corrupt datanode Block. During a write operation, when a datanode writes in to the HDFS, it

verifies a checksum for that data. This checksum helps in verifying the data corruptions during the data transmission.

186. what do you mean by high availability of name node? How is it achieved?

A.) In hadoop version 2.x there are two namenodes one of which is in active state and the other is in passive or standby state at any point of time.

187. Explain counters in MapReduce?

A.) A Counter in MapReduce is a mechanism used for collecting and measuring statistical information about MapReduce jobs and events. Counters keep the track of various job statistics in MapReduce like number of operations occurred and progress of the operation. Counters are used for Problem diagnosis in MapReduce.

188. Why the output of map tasks are spilled to local disk and not in hdfs?

A.) Execution of map tasks results into writing output to a local disk on the respective node and not to HDFS. Reason for choosing local disk over HDFS is, to avoid replication which takes place in case of HDFS store operation. Map output is intermediate output which is processed by reduce tasks to produce the final output.

189. Define Speculative execution?

A.) Speculative execution is an optimization technique where a computer system performs some task that may not be needed. Work is done before it is known whether it is actually needed, so as to prevent a delay that would have to be incurred by doing the work after it is known that it is needed.

190. is it legal to set the number of reducer tasks to zero?

A.) Yes, It is legal to set the number of reduce-tasks to zero if there is no need for a reducer. In this case the outputs of the map task is directly stored into the HDFS which is specified in the setOutputPath

191. where does the data of hive table gets stored?

A.) Hive stores data at the HDFS location /user/hive/warehouse folder if not specified a folder using the LOCATION clause while creating a table.

192. Why hdfs is not used by hive metastore for storage?

A.) Because HDFS is slow, and due to it's distributed and dynamic nature, once something is stored in HDFS, it would be really hard to find it without proper metadata… So the metadata is kept in memory in a special (usually dedicated) server called the namenode ready to be queried.

193. when should we use sort by and order by?

A.) When there is a large dataset then one should go for sort by as in sort by , all the set reducers sort the data internally before clubbing together and that enhances the performance. While in Order by, the performance for the larger dataset reduces as all the data is passed through a single reducer which increases the load and hence takes longer time to execute the query.

194. How hive distribute in the rows into buckets?

A.) Distribute BY clause used on tables present in Hive. Hive uses the columns in Distribute by to distribute the rows among reducers. All Distribute BY columns will go to the same reducer.

195. what do you mean by data locality?

A.) In Hadoop, Data locality is the process of moving the computation close to where the actual data resides on the node, instead of moving large data to computation. This minimizes network congestion and increases the overall throughput of the system.

196. what are the installation modes in Hadoop?

A.) Standalone Mode.

Pseudo-Distributed Mode.

Fully Distributed Mode.

197. what is the role of combiner in hadoop?

A.) Combiner that plays a key role in reducing network congestion. The main job of Combiner a “Mini-Reducer is to handle the output data from the Mapper, before passing it to Reducer.

198. what is the role of partitoner in hadoop?

A.) The Partitioner in MapReduce controls the partitioning of the key of the intermediate mapper output. By hash function, key (or a subset of the key) is used to derive the partition. A total number of partitions depends on the number of reduce task.

199. Difference between Rdbms and noSql?

A.) RDBMS is called relational databases while NoSQL is called a distributed database. They do not have any relations between any of the databases. When RDBMS uses structured data to identify the primary key, there is a proper method in NoSQL to use unstructured data. RDBMS is scalable vertically and NoSQL is scalable horizontally.

200. what is column family?

A.) A column family is a database object that contains columns of related data. It is a tuple (pair) that consists of a key-value pair, where the key is mapped to a value that is a set of columns.

201. How do reducers communicate with each other?

A.) Yes, reducers can communicate with each other by dispatching intermediate key value pairs that get shuffled to another reduce C.

Yes, reducers running on the same machine can communicate with each other through shared memory, but not reducers on different machines.

202. Name the components of spark Ecosystem?

A.) Spark Core

2.Spark SQL

3.Spark Streaming

4.MLlib

5.GraphX

203. what is block report in spark?

A.) Block or report user Block or report isspark. Block user. Prevent this user from interacting with your repositories and sending you notifications.

204. what is distributed cache?

A.) In computing, a distributed cache is an extension of the traditional concept of cache used in a single locale. A distributed cache may span multiple servers so that it can grow in size and in transactional capacity.

205. Normalization vs Denormalization?

A.) Normalization is the process of dividing larger tables in to smaller ones reducing the redundant data, while denormalization is the process of adding redundant data to optimize performance. – Normalization is carried out to prevent databases anomalies.

206. how can you optimize the mapreduce jobs?

A.) Proper configuration of your cluster.

LZO compression usage.

Proper tuning of the number of MapReduce tasks.

Combiner between Mapper and Reducer.

207. what are the advantages of combiner?

A.) Use of combiner reduces the time taken for data transfer between mapper and reducer.

Combiner improves the overall performance of the reducer.

It decreases the amount of data that reducer has to process.

208. what are different schedulers in yarn?

A.) There are three types of schedulers available in YARN: FIFO, Capacity and Fair. FIFO (first in, first out) is the simplest to understand and does not need any configuration. It runs the applications in submission order by placing them in a queue.

209. Explain Hive metastore and Warehouse?

A.) A Hive metastore warehouse (aka spark-warehouse) is the directory where Spark SQL persists tables whereas a Hive metastore (aka metastore_db) is a relational database to manage the metadata of the persistent relational entities, e.g. databases, tables, columns, partitions.

210. Difference between Hive vs beeline?

A.) The primary difference between the Hive CLI & Beeline involves how the clients connect to ApacheHive. The Hive CLI, which connects directly to HDFS and the Hive Metastore, and can be used only on a host with access to those services. Beeline, which connects to HiveServer2 and requires access to only one .jar file: hive-jdbc-version-standalone.jar

211. what are temorary tables in hive?

A.) temporary table is a convenient way for an application to automatically manage intermediate data generated during a complex query. Rather than manually deleting tables needed only as temporary data in a complex query, Hive automatically deletes all temporary tables at the end of the Hive session in which they are created.

212. what is lateral view?

A.) The LATERAL VIEW statement is used with user-defined table generating functions such as EXPLODE() to flatten the map or array type of a column.The explode function can be used on both ARRAY and MAP with LATERAL VIEW.

213. what is the purpose of view in hive?

A.) Views are similar to tables, which are generated based on the requirements. We can save any result set data as a view in Hive ; Usage is similar to as views used in SQL

214. Handling nulls while importing data?

A.) To force Sqoop to leave NULL value blank during import, put the following options in the Sqoop command line: –null-string The string to be written for a null value for string columns. –null-non-string The string to be written for a null value for non-string columns.

215. how is spark better than Hive?

A.) Hive is the best option for performing data analytics on large volumes of data using SQLs. Spark, on the other hand, is the best option for running big data analytics. It provides a faster, more modern alternative to MapReduce.

216. Processing of big tables in spark?

A.) Spark uses SortMerge joins to join large table. It consists of hashing each row on both table and shuffle the rows with the same hash into the same partition. There the keys are sorted on both side and the sortMerge algorithm is applied.

217. Benifits of window function in spark?

A.) Spark Window functions are used to calculate results such as the rank, row number e.t.c over a range of input rows and these are available to you by importing org.apache.spark.sql.functions._

218. Difference between window functions and group by?

A.) GROUP BY functionality only offers aggregate functions; whereas Window functions offer aggregate, ranking, and value functionalities. SQL Window function is efficient and powerful. It not only offers GROUP BY aggregate functionality but advanced analytics with ranking and value options.

219. How can we add a column to dataframe?

A.) use withColumn () transformation function.

220. Difference between logical and physical plan?

A.) Logical Plan just depicts what I expect as output after applying a series of transformations like join, filter, where, groupBy, etc clause on a particular table. Physical Plan is responsible for deciding the type of join, the sequence of the execution of filter, where, groupBy clause, etc. This is how SPARK SQL works internally!

221. Benifits of scala over python?

A.) Scala is a statically typed language that provides an interface to catch the compile time errors. Thus refactoring code in Scala is much easier and ideal than Python. Being a dynamic programming

language, testing process, and its methodologies are much complex in Python

222. How to enforce schema on a data frame?

A.) What Is Schema Enforcement? Schema enforcement, also known as schema validation, is a safeguard in Delta Lake that ensures data quality by rejecting writes to a table that do not match the table's schema.

223. Benifits of enforce schema over default schema?

A.) Because objects are no longer tied to the user creating them, users can now be defined with a default schema. The default schema is the first schema that is searched when resolving unqualified object names.

224. what are the challenges faced in spark?

A.) No space left on device:

This is primarily due to executor memory, try increasing the executor memory. Example --executor-memory 20G

Caused by: org.apache.spark.SparkException: Exception thrown in awaitResult:

The default spark.sql.broadcastTimeout is 300 Timeout in seconds for the broadcast wait time in broadcast joins.

To overcome this problem increase the timeout time as per required example

--conf "spark.sql.broadcastTimeout= 1200"

225.How is scala is different from other languages?

A) Though there are a lot of similarities between the two, there are many more differences between them. Scala, when compared to Java, is relatively a new language. It is a machine-compiled language, whereas Java is object-oriented. Scala has enhanced code readability and conciseness.

226. what is functional programming in scala?

A.) Functional programming is a programming paradigm that uses functions as the central building block of programs.

In functional programming, we strive to use pure functions and immutable values.

227. what is spark config?

A.) SparkConf allows you to configure some of the common properties (e.g. master URL and application name), as well as arbitrary key-value pairs through the set () method.

228. can we configure cpu cores in spark context?

A.) The more cores we have, the more work we can do. In spark, this controls the number of parallel tasks an executor can run. From the driver code, SparkContext connects to cluster manager (standalone/Mesos/YARN).

229. how does partition happen while creating RDD?

A.) In case of compressed file you would get a single partition for a single file (as compressed text files are not splittable). When you call rdd.repartition (x) it would perform a shuffle of the data from N partititons you have in rdd to x partitions you want to have, partitioning would be done on round robin basis.

230. To rename a column in Dataframe to some other name? how to achieve that?

A.) Using Spark withColumn() function we can add , rename , derive, split etc a Dataframe Column.There are many other things which can be achieved using withColumn() which we will check one by one with suitable examples.]

231. Difference between spark 1.6 and 2.x?

A.) Even though Spark is very faster compared to Hadoop, Spark 1.6x has some performance issues which are corrected in Spark 2.x.

they are

sparkSession

Faster analysis

added SQL features

MLib improvements

New streaming module

Unified dataset and data frame API's

232. How do you decide number of executors?

A.) Number of executors is related to the amount of resources, like cores and memory, you have in each worker.

233. How to remove duplicates from an array of elemets?

A.) The ways for removing duplicate elements from the array:

Using extra space

Constant extra space

Using Set

Using Frequency array

Using HashMap

234. what is diamond problem in spark and how to resolve it?

A.) Diamond problem occurs when we use multiple inheritance in programming languages like C++ or Java.

The solution to the diamond problem is default methods and interfaces. We can achieve multiple inheritance by using these two things. The default method is similar to the abstract method. The only difference is that it is defined inside the interfaces with the default implementation.